RAEL: Revista Electrónica de Lingüística Aplicada
Vol./Núm.: 19.1
Enero-diciembre 2020
Páginas: 136-163
Artículo recibido: 27/12/2020
Artículo aceptado: 06/05/2020
Artículo publicado 29/06/2020
Url: http://www.aesla.org.es/ojs/index.php/RAEL/article/view/368

Knowledge-based rules for the extraction of complex, fine-grained locative references from tweets

Reglas basadas en conocimiento para la extracción de referencias locativas complejas en tweets

NICOLÁS JOSÉ FERNÁNDEZ MARTÍNEZ[1]
UNIVERSIDAD CATÓLICA SAN ANTONIO DE MURCIA

CARLOS PERIÑÁN PASCUAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

The automatic analysis of user-generated text content from social media involves the challenge of extracting the locative references mentioned in microtexts, so that their geographic coordinates can be identified and the locations can be pinpointed on a map in geolocation systems. The goal of this article is to describe a knowledge-based model that captures a wide variety of locative references, ranging from geopolitical entities and natural landforms to points of interest and traffic ways, from English and Spanish tweets.

Keywords: *location detection; location extraction; geolocation; named entity recognition*

El análisis automático de texto generado por usuarios de redes sociales supone un reto a la hora de extraer referencias locativas de los microtextos, para la posterior identificación de sus coordenadas geográficas y ubicación en un mapa en sistemas de geolocalización. El propósito de este artículo es describir un modelo basado en conocimiento que es capaz de extraer una amplia variedad de referencias locativas, desde entidades geopolíticas o accidentes geográficos hasta puntos de interés o carreteras, de tweets en inglés y en español.

Palabras clave: *detección de localizaciones; extracción de localizaciones; geolocalización; reconocimiento de entidades nombradas*

1. INTRODUCTION

A locative reference, also called named entity of place, refers to a named space that can be computed by means of latitude-longitude coordinates, among other geographic representations (Purves & Derungs, 2015). These named entities of places have been coined 'toponyms' or 'geographical names' in the linguistic and geographic literature (Quirk et al., 1985; Levinson, 2003; Purves et al., 2018). Location extraction consists in the identification and retrieval of

[1] Corresponding author. Email address: njfernandezmartinez@gmail.com

locative references from natural language texts through either probabilistic-based algorithms or symbolic models that make use of text-mining and rule-based strategies (Middleton et al., 2018; Purves et al., 2018). Indeed, this task corresponds to the area of Geographic Information Retrieval (GIR) (Jones & Purves, 2008), a topical subject that interconnects the fields of Computational Linguistics, Natural Language Processing (NLP), and Artificial Intelligence, *inter alia*. Of particular importance are the practical applications of location detection in social media. For instance, Twitter-based location-extraction systems are particularly useful in disease tracking and health surveillance (Eke, 2011; Dredze et al., 2013), disaster management and tracking (Verma et al., 2011; Crooks et al., 2013; Imran et al., 2014; Jongman et al., 2015; Martínez-Rojas et al., 2018; Zhang et al., 2019) and traffic-incident detection and road traffic control (Gonzalez-Paule et al., 2019), among other uses. However, dealing with tweets proves to be a difficult task, given the linguistic peculiarities of the microtext genre (Baldwin et al., 2013).

In this context, this article addresses the language-based rules that underlie LORE (LOcative Reference Extractor), a rule-based model for the extraction of complex, fine-grained locative references in English and Spanish microtexts, e.g. tweets.[2] When deployed in real-life crisis and emergency scenarios, an NLP system that integrates this model can be of great help for emergency responders, considering the great precision and recall delivered by these rules in the evaluation stage, and the quick performance of the rule-based system. In this respect, LORE outperformed general-purpose off-the-shelf NER tools such as Stanford NER, spaCy, NLTK, and OpenNLP, by achieving a precision score of 0.81, a recall score of 0.81 and an F1 score of 0.81 for the English tweets. For the Spanish tweets, LORE achieved a precision of 0.64, a recall of 0.72 and an F1 score of 0.67. It also showed very quick processing speed, surpassing others, at least in the case of the English model. The remainder of this article is structured as follows. Section 2 introduces the task of location extraction with respect to Computational Linguistics and NLP, describes the tweet genre as a type of microtext with its own linguistic peculiarities, and explores some practical applications that leverage tweets in emergency-based situations. This section also gives insights into the representation of spatial knowledge in natural languages, and a critical review of major research work in location-detection from tweets in recent years. Section 3 provides not only a definition of what we mean by locative references but also an overview of how our language-based rules were built from two development corpora. Section 4 presents the results obtained in the evaluation stage and compares the performance of LORE against general-purpose entity recognizers, and also provides the typology of language-based rules, together with definitions and examples. Section 5 presents some conclusions.

2. BACKGROUND AND RELATED WORK

2.1. *Location extraction, Computational Linguistics, and NLP*

Location extraction or detection, also called ‘georeferencing’ or ‘geoparsing’ in some contexts (Gelernter & Balaji, 2013; Purves et al., 2018), is a task that belongs to the field of Information Extraction and Information Retrieval, and more specifically, GIR (Jones & Purves, 2008). It deals with the identification and retrieval of locative references from natural language texts through probabilistic-based (e.g. Machine Learning and Deep Learning) or symbolic-based methods (e.g. rule-based and lexical approaches) (Middleton et al., 2018; Purves et al., 2018).

[2] LORE, which has been developed in C# with ASP.NET 4.6 and MySQL Database, is freely accessible from the FunGramKB website (http://www.fungramkb.com/nlp.aspx).

Location extraction requires an interdisciplinary approach with the convergence of areas such as Computational Linguistics and NLP (Jones & Purves, 2008; Purves et al., 2018; Yingjie Hu, 2018a). All these research areas deal with unstructured text data and the geospatial information contained therein, which is particularly plentiful in the World Wide Web (Jones & Purves, 2008; Yingjie Hu, 2018b; Purves et al., 2018; Hamzei et al., 2019).

The natural language ambiguity characteristic of unstructured text constitutes in itself a great challenge for GIR systems in the retrieval of locative references, the extraction of spatial relationships, and the location disambiguation process of the extracted spatial knowledge (Frank & Mark, 1991; Al-Olimat et al., 2019). Natural language ambiguity is further exacerbated by the noisy, informal and abbreviated nature of the microtext genre (Baldwin et al., 2013; Eisenstein, 2013). In this regard, expertise in Linguistics and Computational Linguistics becomes of utmost importance for the extraction and representation of locative references and spatial expressions in a structured, digitalized format (Stock et al., 2019).

2.2. *The tweet genre and practical applications of Twitter-based geolocation systems*

Among social media and, in particular, among microblogging services, Twitter stands out as one of the most popular worldwide microblogging platforms for information sharing and communication purposes (Murthy, 2018). In Twitter, users can post microtexts (i.e. tweets) which are brief, character-limited messages (280 characters maximum) that typically express the users' thoughts, activities, and opinions about their daily lives or about a given topic (Hu et al., 2013).

Microtexts are usually informal, noisy and abbreviated, so language conventions generally deviate from the linguistic norm through language devices such as abbreviations (e.g. *pls* instead of *please*), acronyms (e.g. *FYI* instead of the phrase *for your information*), misspellings (e.g. *madrizz* instead of *Madrid*), lack of capitalization (*united kingdom* instead of *United Kingdom*), ungrammatical forms (e.g. *you was* instead of *you were*), ellipsis and truncated sentences (e.g. *incident in Newcastle* instead of *There was an incident in Newcastle*) (Baldwin et al., 2013; Eisenstein, 2013). In this regard, one particular challenge in the identification of locative references in tweets is related to the linguistic peculiarities of the microtext genre. Most NLP systems, which have historically been trained on formal written texts, such as those from the news genre or the scientific literature, face unexpected problems when applied to tweets, which is why their performance is usually much degraded (Hoang & Mothe, 2018). These systems have been designed to rely on proper spelling, capitalization and grammatical patterns for different NLP tasks, e.g. part-of-speech (POS) tagging or chunking; consequently, in the absence of these linguistic conditions, their predictive power decreases. Several strategies have been proposed to overcome the present linguistic difficulties in NLP systems applied to Twitter, such as the normalization of the tweet text (Liu et al., 2012) and/or the adaptation of NLP tools to social-media genres and their linguistic idiosyncrasies (Eisenstein, 2013). However, despite the widely-believed claim that tweets are noisy and informal, Hu et al. (2013) disagree as to the apparent degree of informality of the tweet genre, arguing that tweets are not as informal as other microtext genres (e.g. SMS). In fact, according to the authors, tweets can be considered as a projection of other formal textual genres onto a size-restricted format.

Geolocation systems play a key role in a variety of real-life scenarios, particularly when geospatial information proves vital to allocate resources and services to affected areas and persons in times of crisis and emergencies (Martínez-Rojas et al., 2018). For instance, in health-related scenarios such as health surveillance or disease tracking, geospatial information obtained from social-media microtexts can be exploited by public health and medical officials for tracking or prevention measures in disease propagation or forecasting (Eke, 2011; Dredze

et al., 2013; Vilain et al., 2019; Singh et al., 2020). Many emergency-based services employ natural or human-made disaster detection and tracking systems with a geolocation module in floods, earthquakes, storms, civil unrest, war, crime, etc. (Vieweg et al., 2010; Crooks et al., 2013; Imran et al., 2014; Jongman et al., 2015; Martínez-Rojas et al., 2018; Zhang et al., 2019). Geolocation systems can also be vital for traffic incident detection and road traffic control (Ahmed et al., 2019; Gonzalez-Paule et al., 2019; Khodabandeh-Shahraki et al., 2019). Another practical application derived from geolocation systems is that of marketing and advertising, where the locations mentioned in tweets can be exploited to suggest potential places for Twitter users to visit (Li & Sun, 2014).

2.3. *Location-detection systems for tweets*

In recent years, many systems have been proposed to extract locative references from tweets, with more or less success. A critical review of major works in the field of location extraction in tweets is presented in chronological order in this section.

Gelernter & Balaji (2013) proposed an algorithm for local microtext geoparsing using regex-based rules, the Open Calais Named Entity Recognition (NER) software, machine-learning techniques for abbreviation disambiguation, and a geodatabase with place names from New Zealand and Australia at and within city level, such as geopolitical entities, points of interest (POIs), buildings, and streets. For the evaluation stage, they used a corpus of tweets about the 2011 Christchurch earthquake in New Zealand, achieving an F1 score of 0.9. The model was also tested on an evaluation corpus of tweets about the 2011 Texas wildfire in the US, obtaining an F1 score of 0.71. The reason why this algorithm achieved a very high F1 score for the first dataset is explained by the fact that the training corpus and evaluation corpus shared the same emergency event, and the algorithm might have been particularly skewed for the location types mentioned in the Christchurch earthquake. The F1 score of the Texas fire evidenced that in other events the algorithm may suffer from poorer performance. In this regard, it has been reported that case studies of particular disaster events with well delimited spatial boundaries usually yield higher evaluation results (Karimzadeh et al., 2019). It would thus remain to be seen whether such high performance could be replicated with global-scale events or local events other than those that may occur in New Zealand or Australia. Another downside of this study has to do with the compilation of an ad-hoc location-indicative noun dataset, with only a few building and address types.

Malmasi & Dras (2016) proposed a linguistic-based unsupervised location-detection model based on linguistic techniques and rules such as NP extraction and n-gram matching techniques using regex rules and the GeoNames database (Ahlers, 2013). It targeted geopolitical entities, POIs, buildings, addresses, and surrounding distance and direction markers, giving an F1 score of 0.792. This research provided a more linguistic-based focus for the task of location detection. However, there are some drawbacks that need to be discussed, e.g. (a) the debatable rigor in the authors' decision to create ad-hoc lists of location indicative words (addresses, POIS…), and (b) a loose evaluation metric standard performed on a per-token basis, instead of per-location entity, both of which might have contributed to a higher F1 score.

Middleton et al. (2018) proposed several location-detection and location-disambiguation models, of which the best was a location-detection model for English tweets using the OpenStreetMaps database (Acheson et al., 2017). It used NLP techniques such as a sentence tokenizer, n-grams for matching tokens against the OpenStreetMaps database, and their own corpus of building and street types, among other lexical resources. They focused on geopolitical entities, buildings, and streets. Their training dataset of tweets was imported from different news events. The evaluation stage was carried out for separate corpora of tweets about different

incidents (i.e. blackout, earthquake, and hurricane) in different geographic areas (i.e. Christchurch, Milan, New York, and Turkey) for which the geodatabase was preloaded with locations for those specific areas for the evaluation of each corpus. F1 scores ranged from 0.90 to 0.97. A disadvantage of their model is the very slow processing speed, since it has to preload many locations in memory before deploying the location-extraction module. Overall, the authors highlighted the importance of implementing linguistic knowledge and using geodatabases in location extraction from tweets to achieve great results. It would be interesting to see whether the application of their model to global-scale corpora of tweets about different issues and targeting more location types delivers the same results, and how processing speed becomes affected.

Al-Olimat et al. (2018) proposed an unsupervised location-detection model for tweet texts by leveraging the GeoNames database with an n-gram model complemented by collocational information. It was applied on three tweet datasets corresponding to local flood events in Chennai, Louisiana and Houston respectively, achieving an F1 score of 0.81 on a per-token evaluation basis. However good the results are, we are not provided with an explanation about the location types extracted by their model.

Dutt et al. (2018) developed an unsupervised location-detection model for tweets based on regex-based rules, ad-hoc lists of location-indicative words, syntactic chunking and dependency parsing, the Spacy NER tagger, and GeoNames. The system achieved an F1 score of 0.81 on a per-entity-based evaluation. It was applied to a large test corpus of tweets (239,256 tweets) collected using the keywords *dengue* and *flood* for emergency-related events of those types located in India. The methodology followed is linguistic-based in that they made use of linguistic knowledge and NLP techniques for NER. The authors did not present information about the location types extracted by their model.

Hernandez-Suarez et al. (2019) proposed a NER-based system for detecting and geocoding toponyms (e.g. street, avenue, building, region or country) in Spanish tweets about the 2017 Mexico City earthquake. The system was grounded on a deep-learning model and pre-trained word embeddings using the corpus of Spanish tweets as training data. Their algorithm achieved an F1 score of 0.80.

Singh et al. (2020) provided an in-depth study of the current coronavirus COVID-19 pandemic, in which they also focused on locative references mentioned in tweets dealing with the COVID-19 outbreak. They used Wikipedia and Statoids, two major databases to extract geopolitical entities such as countries, states, provinces, and cities. With this geospatial information, they analyzed the correlation between the number of confirmed cases in different regions of the world and the number of location mentions in the tweets, finding a high correlation between both: the more confirmed cases of coronavirus in a given area, the more that area appeared mentioned in the tweets. Singh et al. (2020) underlined the importance of location extraction techniques to study the evolution and spread of pandemics and for disease forecasting.

Wang et al. (2020) built a location extractor called NeuroTPR using a deep-learning framework with rich linguistic-based features for the task of location extraction from tweets. For the training phase, they employed 599 tweets, together with automatically annotated location-related chunks from the Wikipedia. For the evaluation of their tool, they used different corpora: (a) a tweet corpus about Hurricane Harvey, (b) GeoCorpora (Wallgrün et al., 2018), which is also made up of tweets, and (c) a dataset with chunks from the Web. They focused on location types such as geopolitical entities, natural landforms, POIs, and a few traffic ways. In the evaluation stage, they compared their model against standard off-the-shelf NER tools such as Stanford NER, spaCy, and other deep-learning models retrained with their training dataset. NeuroTPR achieved a precision score of 0.787, a recall score of 0.678, and an F1 score of 0.728 on the Hurricane Harvey corpus. Stanford NER showed great precision numbers (i.e. 0.828),

but the other deep-learning models achieved evaluation numbers lower than those of NeuroTRP, where spaCy obtained much worse results (i.e. F1 score of 0.366). With GeoCorpora, NeuroTPR achieved a precision score of 0.8, a recall score of 0.761, and an F1 score of 0.78.

2.4. *The representation of spatial knowledge in natural languages*

According to the linguistic literature (Herskovits, 1985; Landau & Jackendoff, 1993; Talmy, 2000; Kracht, 2002; Levinson, 2003; Coventry & Garrod, 2004; Bennett & Agarwal, 2007; Radke et al., 2019; Stock et al., 2019), spatial knowledge, typically represented by spatial prepositions in analytical languages, indicates a spatial relationship held by different entity types or arguments, formally expressed as S(*x*, *y*), where (a) S determines the kind of spatial relationship held by *x* and *y*, (b) *x* refers to what is spatially defined, and (c) *y* represents the region of space occupied by *x*.

2.4.1. *The structural and syntactic features of spatial expressions*

From a structural standpoint, in Western European languages such as English, Spanish, French, or Italian, a spatial expression is generally composed of a 'subject' (i.e. what is located) and a prepositional phrase (PP) made up of a preposition and an 'object' (i.e. where is located). This PP can modify a noun (e.g., *the glass on the table*), or predicate something about a noun phrase (NP) (e.g. *John is at school*) or a clause (e.g. *He is buying groceries at the market*) (Geis, 1975; Herskovits, 1985; Creary et al., 1989). The object refers to a physical location, real or imaginary, which describes the position, direction/path, or distance of a given entity (Kracht, 2002; Coventry & Garrod, 2004; Bennett & Agarwal, 2007; Cinque & Rizzi, 2010). Whereas position indicates a spatial relationship of location among objects, path specifies a trajectory understood in terms of source and goal, and distance provides a measure of space among two or more entities.

In natural languages, places are typically encoded as nouns, which can be proper if used to identify a specific and unambiguous spatial region or portion (e.g. *Granada*, *Valencia*, *Spain*, *France*), receiving the name of 'toponym' or 'place name' (Levinson, 2003; Stock et al., 2019), or common when they are used in a generic sense, often representing a semantic type of different granularity (e.g. *neighborhood, city, country, beach, canyon, street, road*). They can also be formally represented by means of complex NPs (e.g. *the black chair next to the table standing in the corner*), which can recursively become very intricate, especially if multiple reference frames are mentioned (Stock et al., 2019).

As far as syntax is concerned, spatial expressions can be found at the beginning (e.g. *In Tokyo the earthquake caused great damage*) or at the end of the clause (e.g. *Floodings were reported in New Jersey*). As mentioned above, these expressions can specify different semantics, such as position (e.g. *John lives in New York*), direction (e.g. *An ambulance is heading to Glenwood Avenue*), or distance (e.g. *Mary drove for 35 miles southwest of London*) (Quirk et al., 1985: ch. 8). Their referents can be the subject (e.g. *Paul flew to Los Angeles*), the direct object (e.g. *I parked the car at Nevada Shopping*), or even both (e.g. *I met Anna at the National Museum*). Spatial expressions typically perform the adverbial function in the clause (Geis, 1975), although they can also act as postmodifiers of a noun in a NP when formally realized as PP (e.g. *The man outside the bus station is waiting for his friends*) (Quirk et al., 1985). According to Quirk et al. (1985), spatial expressions performing the adverbial syntactic function of space adjuncts can be either obligatory (e.g. **John lives*) or optional (e.g. *We bought groceries* (*at Tesco))* (Geis, 1975). When obligatory, these syntactic units additionally perform the function of postmodifiers with verbs of stative meaning (e.g. *be, live, stand, lie*, etc.). The formal realization of these phrases as space adjuncts can be NP (e.g. *John*

walked five miles), PP (e.g. *Mary was a teacher in Newcastle*), Adverbial Phrase (AP) (e.g. *The warriors died there*) or subordinate clauses of distinct complexity (e.g. *The missing boy was found where the police could have not ever imagined*). PP is the most typical phrasal realization and the most connected with spatial expressions (Quirk et al., 1985: Ch. 9). Since our interest is in place names, and these are nouns, we only take into account NPs and PPs.

Spatial prepositions act as linkers to encode spatial relations between objects or between an object and a region/place (Landau & Jackendoff, 1993). A distinction should be made between those spatial prepositions indicating location (i.e. locative prepositions *in, at, near*, etc.) and spatial prepositions indicating direction (i.e. directional prepositions such as *to*/*from*) (Coventry & Garrod, 2004). Locative prepositions can be further divided into topological terms that express topological relations among entities (e.g., *in, at, on, near*, etc.) and projective terms that need a frame of reference (e.g., *in front of, above, to the right*, etc.). In this context, the prepositions *in* and *at* are prototypical items of locative prepositions (Levinson, 2003). These obey different patterns for their usage in discourse: *in* is usually reserved for large geopolitical entities such as districts, regions, cities, countries, continents, etc., or to refer to the dimensional side of buildings (e.g. *John works in a record company*), whereas *at* is rather used with small geopolitical entities (e.g. *Mary lives at Stratford-upon-Avon*) and buildings in the institutional and functional sense (e.g. *John works at a record company*) (Quirk et al., 1985; Vasardani et al., 2013).

2.4.2. *Named entities of places: Toponyms and geographical names*

On the one hand, toponyms or place names can be defined as named place specifications that, from a geographical and mathematical point of view, do not provide by themselves a precise frame of reference, which is typical of quantitative methods involving coordinate systems (Levinson, 2003). They can be accordingly cast into a generic semantic class (e.g. London:city) (Bennett & Agarwal, 2007). Ascribing a place name to a particular location is a special type of topological relation whereby the place name acts as the ground location of a given figure (e.g. *John lives in London*) (Levinson, 2003). In this sense, as Levinson (2003: 69) claims, toponyms offer an "underlying mental map of locations" which speakers can have access to and more or less place on a map.

Geographical names include place names in their lexical scope with the addition of location-indicative nouns, also called "descriptors" with an "appositive function" (Quirk et al., 1985: 1317): e.g. *Mount Everest, New York State, Sunset Boulevard*, etc. In the English language, the 'name-first construction' is especially common, where location-indicative nouns typically follow toponyms (e.g. *Nile Valley, Quebec Province*). It is not rare, however, to find examples of location-indicative nouns preceding place names (e.g. *River Thames*). At times, both can be reversed (e.g. *Cork County* or *County Cork*). At other times, location-indicative nouns and place names can be linked by the preposition *of* as in the *State of Missouri*, the *Island of Cyprus*, or the *coast of New Zealand*.

Toponyms and geographical names alike are often preceded by spatial prepositions (Al-Olimat et al., 2019), though they do not necessarily need to be accompanied by them (e.g. *Madrid is the capital of Spain*). Their extraction in such contexts becomes harder, requiring other location-signaling clues, such as the presence of location-indicative nouns or locative markers, or using a geographic database such as GeoNames in order to match the entities of the text against the database in Named Entity Matching approaches (Middleton et al., 2018).

3. METHODOLOGY

3.1. *Definition of locative reference, location-indicative noun and locative marker*

We define a locative reference as a subtype of named entity that designates a specific, unambiguous, and precise physically-locatable geographic reference i.e. one that can be typically rendered into geographic coordinates or other geospatial measurements and thus pinpointed on a map (Leidner & Lieberman, 2011; F. Liu et al., 2014; Gritta et al., 2018). In linguistic terms, locative references are typically proper nouns that designate named entities of place (i.e. toponyms or geographical names). With respect to their morphology, locative references can be realized as full words (e.g. *Madrid, city of London*), abbreviations (e.g. *FR, bcn*), acronyms (e.g. *UK, US*), alphanumeric codes (e.g. *M-40*), or as a combination of them (e.g. *I-90 SW*). As for their semantics, we establish five main locative categories: geopolitical entities (e.g. *New York*), natural landforms (e.g. *Mount Everest*), POIs (e.g. *Victoria Coach Station*), and traffic ways (e.g. *110 Croydon Road, I-290*). From a structural point of view, we distinguish between simple and complex locative references according to the number and complexity of lexical units that make up one locative reference. In this respect, a simple locative reference is composed of one or several proper nouns (e.g. toponyms such as *Granada, United Kingdom*), whereas a complex locative reference offers a rich lexical network by means of the juxtaposition of location-indicative words and locative markers to the proper noun (e.g. geographical names such as *New York City, Lake Michigan, Borough of Manhattan* or *25 miles NW of London*). Taking into consideration the surrounding lexical elements that comprise locative references in the form of location-indicative nouns and/or locative markers offers more detailed geospatial information (Van et al., 2013), which could ultimately be more useful for competent authorities to trace the location of a given emergency. To illustrate the complexity of locative references, Figure 1 presents the phrasal structure of English locative references, where the asterisk is used to mark optionality, and double asterisk refers to the optional presence of locative markers either at the beginning or at the end of the locative reference.

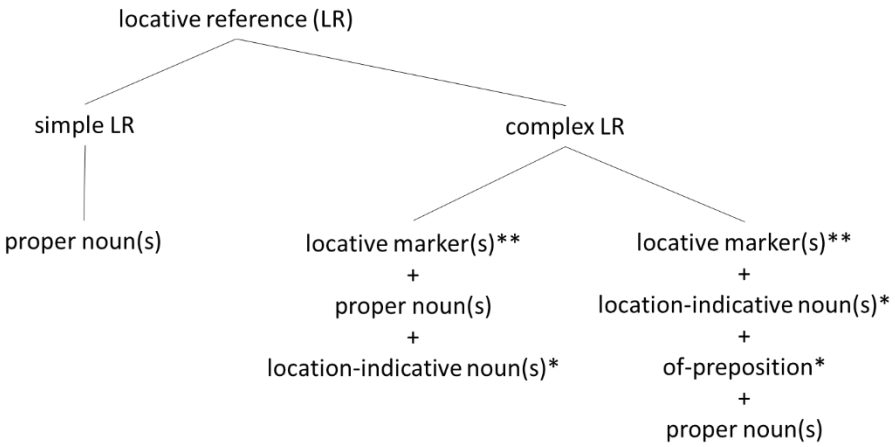


Figure 1: ***The phrasal structure of English locative references***

A location-indicative noun is a common noun that designates a generic place and that may accompany proper nouns in locative references (e.g. *Glenwood Avenue, city of Barcelona, río Guadalquivir*). These nouns were automatically collected from the EuroWordNet lexicon through the synsets "road.n.01", "building.n.01", "facility.n.01", "junction.n.01", "district.n.01", "area.n.01", "geological_formation.n.01", "body_of_water.n.01", "tract.n.01", "way.n.06", and "beach.n.01" to obtain datasets of such nouns in English, Spanish, French, and Italian.[3] Later, a semi-automatic filtering process was applied to discard words whose locative

[3] At the present moment, LORE provides full support for English and Spanish, while support for French and Italian is under development.

meaning is not self-evident (e.g. *bed, melting pot, scene of action, junk pile, parts,* etc.) and more-than-two-word lexical items, because of their rare presence in tweets. Abbreviations were also added to the location-indicative noun dataset, retrieved from a list of traffic-way and other place abbreviations obtained from the US postal service database for the English dataset, and from other sources for the Spanish dataset.[4] Examples of English location-indicative nouns are *country, state, region, province, city, town, kingdom, villa, bay, mountain, mount, ridge, volcano, valley, lake, river, shore, beach, park, canyon, building, museum, school, station, stadium, garden, café, tavern, hospital, court, theater, residence, zoo, casino, square, street, st, boulevard, blvd, avenue, av, alley, road, rd, highway, hwy, freeway, fwy, turnpike, tpk, route,* etc. Examples of Spanish location-indicative nouns include *barrio, academia, ciudad, cima, albergue, autovía, condado, cuenca fluvial, biblioteca, avenida, distrito, desierto, centro médico, calle, isla, cine, camino, lago, litoral, hospital, carril, país, llanura, museo, provincia, montaña, restaurante, parada, urbanización, teatro, río, vía,* etc.

A locative marker specifies a distance in measurable terms (i.e. amount of space or of time), and/or a direction (i.e. path specified by a cardinal point). Examples include *north of New York State, 50 miles SW of Liverpool, 25mins away from Northumbria Street, 25 minutos de la avenida de Madrid,* and *40 kms al noroeste de Bilbao*.

3.2. *The development phase of the language-based rules in LORE*

3.2.1. *Development corpus and evaluation corpus*

In LORE, we employed development corpora of English and Spanish tweets for the study and extraction of full-fledged linguistic patterns, and then we tested the resulting rules with the evaluation corpora. To this end, we compiled our own corpora through an automatic search of tweets, using seven keywords related to crisis and emergency events (i.e. *earthquake*, *flood*, *car accident*, *bombing attack*, *shooting attack*, *terrorist attack*, and *incident*), so that we could retrieve tweets mentioning issues of different nature. Similarly, we used their near-equivalents in Spanish (i.e. *terremoto, inundación, accidente de coche, ataque terrorista, bombardeo, tiroteo* and *incidente*) for the construction of the Spanish corpus. Moreover, we strictly followed corpus linguistic principles in the compilation phase as to representativeness and coverage (Reppen, 2010), meaning that the corpora were large and representative enough for the issue of location extraction. We compiled the English development corpus and evaluation corpus, containing 500 and 800 tweets each, on 8 April 2019 and 11 April 2019, respectively, and the Spanish development corpus and evaluation corpus, containing 100 and 500 tweets each, on 28 May 2019 and 27 August 2019, respectively.

3.2.2. *Development and refinement process of language-based rules*

We carried out the extraction of linguistic patterns by paying special attention to the linguistic idiosyncrasies of the tweet genre and the geospatial features of natural languages, as discussed in Section 2.2 and Section 2.3. In other words, we thoroughly analyzed the different combination of n-grams and the POS of tokens, the presence of locative-contextual clues such as locative prepositions, location-indicative nouns, and locative markers, which usually signal the presence of locative references. All this knowledge was integrated in the formulation of regular expressions that took into consideration the above linguistic variables. Through engaging in continuous evaluations in an 'iterative refinement process' of our rule-based approach (Barrière, 2016), the regular expressions had to be tweaked and fine-tuned to tackle

[4] http://cool.conservation-us.org/lex/abbr_suf.html, http://www.wikilengua.org/index.php/Lista_de_abreviaturas_de_v%C3%ADas, and https://www.abreviaciones.es/edificios-lugares-y-negocios/

natural language ambiguity and the noisy nature of tweets, up to their current high-performance state. This involved looking at error-prone patterns derived from poorly defined regex-based rules, rather than individual errors or missed locative references, which could potentially lead to overfitting and ad-hoc decisions. Our goal was to anticipate and prevent the erratic behavior of the model when applied to any other corpus of tweets. Obviously, this process resulted in rules that were more restrictive than those elaborated at the initial stages. Although each language expresses locative relations in slightly different ways, we used a single inventory of extracted linguistic patterns and rules with the languages supported by our model, i.e. English and Spanish, with a view to extending support to French and Italian.

The multilingual adaptation of LORE to languages other than English did not start from scratch, and only needed a few tweaks in the regex-based rules plus semi-automatic methods for the retrieval of lexical resources. These tweaks and modifications involved taking into account the linguistic peculiarities of Romance languages, which express spatial relations in different ways. For instance, Spanish geographical names start with the location-indicative noun(s) and may incorporate different combinations of prepositions and determiners before arriving at the toponymic part (e.g. *Avenida de la Constitución*). Also, Spanish locative marker constructions are different from the English ones. For instance, complex locative marker constructions have a different lexico-grammatical profile (e.g. *XX mins away from* vs *XX mins hasta/de*). Since these phrasal structures are also found in French and Italian and grammatically encoded in the same way as in Spanish, multilingual support is being extended to these languages using the same built-in regex-based rules. Thus, this resulted in the creation of two types of rules: (a) language-independent rules, which apply to all the languages that are going to be supported by LORE, and (b) language-dependent rules, which were modeled according to the idiosyncrasies of each language.

4. RESULTS AND DISCUSSION

The evaluation of LORE with the English and Spanish corpora followed an entity-based criterion (Gritta et al., 2018; Das & Purves, 2019), i.e. considering whether or not the locative references that were extracted exactly matched those annotated in the evaluation corpora. We also employed a token-based evaluation criterion, i.e. considering partial matches. The evaluation metrics were precision, recall, and F1 (i.e. the harmonic mean of precision and recall), as shown in Figure 2.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Figure 2: ***Commonest evaluation metrics for NER***

True Positive (TP) refers to a correctly identified locative reference. False Positive (FP) refers to instances that have been wrongly identified as locative references. False Negative (FN) is the label used for those locative references that were not captured by the model.

In both the entity-based and token-based evaluation criteria, exact matches count as TP. However, they differ in the treatment of cases of partial or inexact matches. In entity-based evaluation, partial matches can penalize, since they count either as FP when the boundaries of the extracted instance exceed the boundaries of the locative reference (e.g. ~~*Off*~~ *East Coast of Honsu* instead *of East Coast of Honsu*) or as FN when the boundaries of the extracted instance fall short (e.g. *Camino* instead of *Camino Pablo*). In token-based evaluation, partial matches of the type commented above also count as TP, apart from being FP or FN. Thus, entity-based evaluation is the strictest evaluation method. On the other hand, token-based evaluation works more leniently, yielding higher scores.

With the English evaluation corpus, considering the entity-based evaluation criterion, the system achieved a precision of 0.81, a recall of 0.81 and an F1 score of 0.81. With the Spanish evaluation corpus, it achieved a precision of 0.64, a recall of 0.72 and an F1 score of 0.67. The performance of LORE was also compared with well-known off-the-shelf NER tools, such as spaCy (Honnibal & Johnson, 2015), NLTK (Bird, 2006), Stanford NER (Finkel et al., 2005) and OpenNLP (Ingersoll et al., 2013). LORE outperformed all of them in all scores (Table 1 and 2) and, in the case of English, also in processing speed (Table 3 and 4). In bold, we highlight the best scores.

Table 1: ***Evaluation metrics for the English evaluation corpus***

English location-detection model	**Token-based evaluation**			**Entity-based evaluation**		
	P	**R**	**F1**	**P**	**R**	**F1**
LORE	0.85	**0.83**	**0.84**	**0.81**	**0.81**	**0.81**
Stanford NER	**0.89**	0.42	0.57	0.79	0.37	0.50
NLTK	0.55	0.29	0.38	0.43	0.24	0.31
spaCy	0.75	0.33	0.46	0.66	0.28	0.39
OpenNLP	0.73	0.27	0.40	0.56	0.21	0.30

Table 2: ***Evaluation metrics for the Spanish evaluation corpus***

Spanish location-detection model	**Token-based evaluation**			**Entity-based evaluation**		
	P	**R**	**F1**	**P**	**R**	**F1**
LORE	0.73	**0.74**	**0.74**	**0.64**	**0.72**	**0.67**
Stanford NER	**0.87**	0.49	0.63	0.62	0.37	0.48
NLTK	0.33	0.27	0.30	0.23	0.21	0.22
spaCy	0.71	0.62	0.66	0.58	0.55	0.57
OpenNLP	n/a	n/a	n/a	n/a	n/a	n/a

Table 3: ***Processing speed for the English evaluation corpus***

Location-detection model	**Processing speed (min:sec.cs)**
LORE	**00:08.69**
Stanford NER	00:09.82
NLTK	00:10.88
spaCy	00:12.15
OpenNLP	03:35.10

Table 4: ***Processing speed for the Spanish evaluation corpus***

Location-detection model	**Processing speed (min:sec.cs)**
LORE	00:56.75
Stanford NER	00:06.41
NLTK	**00:06.37**
spaCy	00:31.16
OpenNLP	n/a

4.1. *A typology of the linguistic rules*

We present a typology of the regex-based rules that exploit linguistic knowledge and contextual evidence for their application on any of the languages supported in LORE. We tested these rules with the English and Spanish evaluation corpora, showing their effectiveness in the task of extracting locative references from tweets. We provide examples from the evaluation corpora to illustrate the strengths and weaknesses of our rules. Whenever the rules failed in the extraction of locative references, we provide an explanation to account for their faulty behavior, and suggest potential solutions for a future refinement process.[5]

4.1.1. *Rules for n-gram combinations of locative references using a geodatabase*

These rules are language-independent and apply to n-grams, i.e. linear sequences of *n* words in a given sample of text. To understand n-grams, let us consider the sentence *The quick brown fox jumps over the lazy dog*. In it, we can find unigrams or n-grams of size n = 1 (e.g. {the}, {quick}, {brown}, {fox}...), bigrams or n-grams of size n = 2 (e.g. {the quick}, {quick brown}, {brown fox}…), trigrams or n-grams of size n = 3 (e.g. {the quick brown}, {quick brown fox}…), and so on.

In particular, the rules deal with bigrams and unigrams when matching the tokens in the tweets against the tokens found in our place-name dataset built from GeoNames (cf. Flowchart #1 in Appendix 1):

i) For bigrams, if (a) the first token is not a noun, and (b) the second token is not a proper noun, or the second token is a directional marker (e.g. *South*, *sur*), it is very likely that the n-gram is not a locative reference. Examples of bigrams taken from the corpora that were found in GeoNames but are not actual locative references according to the linguistic context are *the country, beautiful isle, nice airport, the South, el tiroteo* ('the shooting'), *la bomba* ('the bomb'), *buenas tardes* ('good evening'), *de armas* ('of weapons'), *el Sur* ('the South'), etc.

ii) For unigrams, (a) if the unigram is not a proper noun, or (b) if the unigram is in the stopword dataset,[6] the location-indicative noun dataset or the locative marker dataset, it is very likely that it is not a locative reference. Examples of unigrams taken from the

[5] Rules are indicated by the numeration (i), (ii), (iii), etc. Examples are offered using cardinal numbers (1), (2), (3), etc. For exceptions to or specific conditions derived from the application of the rules, we use the letters (a), (b), (c), etc.

[6] A stopword dataset is commonly used in NLP systems to filter and discard very frequent words that might compromise the precision of the model. We used the 5000 most frequent words from the COCA corpus (https://www.wordfrequency.info/) together with a list of person names and surnames from https://names.mongabay.com/ and https://surname.sofeminine.co.uk/w/surnames/most-common-surnames-in-great-britain.html. For Spanish, we used the 20000 most frequent words from the *Corpus del Español* (https://www.wordfrequency.info/files/spanish/spanish_lemmas20k.txt) and a name and surname list from https://github.com/olea/lemarios.

corpora that were found in GeoNames but are not actual locative references according to the linguistic context are *police, going, Ashley, accident, Clinton, Gracias* ('thanks'), *terremoto* ('earthquake'), *camping, compartir* ('share'), *López, amor* ('love'), *coche* ('car'), etc.

At times, these rules were bypassed by certain n-grams that were captured in GeoNames but that were unfortunately not filtered by the stopword dataset, especially in the case of person names (e.g. *Jam*, *Yao*, *Robles*, *Nemo*, *Obama*, etc.). The rules and datasets thus play a preventive role which might not always avoid the extraction of wrong instances, since these person names can also be actual locations and only the linguistic context can disambiguate them. Therefore, we searched for a trade-off between precision and recall when using these rules and datasets.

4.1.2. *Rules that exploit locative prepositions*

The following rules are language-independent. If a token is a locative preposition, then it is very likely that any succeeding combination of proper nouns is a locative reference, except when (a) the proper noun is a date, or (b) the proper noun is a person name or any other type of named entity, ruled out by the stopword dataset (cf. Flowchart #2 in Appendix 1). The following examples of actual tweets from the corpora illustrate the extracted locative references. Example (1) shows the extracted unigram *Palakkad*, thanks to the presence of the locative preposition *at*.

(1) *Visited home of Mr. Shobha Aboobacker Sahib at Palakkad who passed away today morning in an accident*[7]

In and *across* are other prepositions that can signal a locative reference, illustrated by the examples below:

(2) *When you're doing your show in San Bernardino...and you need a listener to tell you about a 3.5 earthquake*
(3) *Golestan province N Iran Three weeks after the floods, the houses are still surrounded by floods in Aqqala.*
(4) *Floods in #Iran - Villages in #Khuzestan surrounded by floods, no sign of state relief. #IranFloods #IranRegimeChange.*
(5) *#GhassemSoleymani very clearly doesn't care about #flood and its victims across Iran.*

In Spanish, *en* is the most prototypical locative preposition, as shown in the following examples:

(6) *La #Tormenta en MADRID pone de manifiesto, otra vez, el lamentable estado de las infraestructuras*
'The storm in Madrid exposes, once again, the lame conditions of the infrastructures'
(7) *Vuelve a caer más fuerte que antes en Valdemoro, ahora con aparato eléctrico.*
'It rains more heavily than before in Valdemoro, now with thunder and lightning'

[7] Before any location-extraction task, tweet preprocessing is performed whereby user mentions and URLs are replaced by the tokens "user" and "url", respectively. Moreover, emojis and other special characters are removed, as well as extra white spaces, and words in hashtags are segmented. No normalization techniques are applied for spelling.

(8) *Inundaciones en Arturo Soria. Garajes inundados, ahora cae piedra #Madrid @112cmadrid @E112Andalucia.*
'Floods in Arturo Soria. Flooded garages, dropping stones now #Madrid @112cmadrid @E112Andalucia'

(9) *#NuevoLeon Fuertes lluvias en Nuevo León dejan dos muertos e inundaciones*
'#NuevoLeon Heavy rains in Nuevo León kill two people and cause floods'

(10) *Agresiones al ejército en Michoacán - Severos daños por lluvias en Sinaloa*
'Assaults on the army in Michoacán – Serious damage caused by rains in Sinaloa'

Now we present other tweets in which our rules and datasets did not manage to detect the locative references. In Example (11), *Indinapuram* was missed because *between* was not considered a locative preposition in the English lexical dataset due to its ambiguity in some contexts and its less-than-prototypical spatial nature.[8] In this regard, we also excluded the directional prepositions *to* and *from*, considering the cost-benefit ratio of their ambiguous nature, since they appear with ditransitive constructions (e.g. *give, obtain, receive,* etc.) typically followed by person names.

(11) *Pls consider asking the #NHAI to close the central verge on #NH24 between #Indirapuram and…*

Rules were constructed with respect to the languages supported by LORE. Therefore, only proper nouns that follow locative prepositions are considered, so the rules cannot for now handle the combinations of proper nouns with words of different grammatical categories, e.g. determiners, prepositions, etc., as shown in Example (12).

(12) *Se desborda rio en Los Reyes*
'Overflowed river in Los Reyes'

At other times, n-gram combinations were wrongly detected as locative references. In Example (13), *Mandarin* was extracted as a locative reference because, according to the POS tagger, its grammatical category is proper noun. Since it was preceded by the preposition *en*, and the stopword dataset could not filter it out, it was wrongly retrieved as a locative reference.

(13) *Si claro como no...ahora digame el chiste en ~~Mandarin~~ por favor!!*
'Yeah, yeah, of course…now tell me the joke in Mandarin, please!!'

4.1.3. *Rules that exploit location-indicative nouns*

These rules are language-dependent. On the one hand, in the case of English, there are several cases in which a combination of tokens including a location-indicative noun refers to a locative reference (cf. Flowchart #3 in Appendix 1). For example:

i) when location-indicative nouns are preceded by one or a combination of proper nouns

(14) *Pattonville Fire Protection District is currently responding to an emergency incident for a(n) 13 Diabetic Problems QD*

[8] Ambiguity from an NLP perspective refers to the inability of machines to disambiguate more than one meaning. This phenomenon is more commonly known as 'polysemy' in Theoretical Linguistics.

(15) *Westville Public Schools is having a mock accident today at 10 am. Please do not be alarmed at all of the EMS*
(16) *Incident on #LLine Both directions from Myrtle Avenue Station to Rockaway Parkway-Canarsie Station*
(17) *Rising Seas May Mean Tampa Bay Floods Even During Sunny Days*

ii) when one of the preceding tokens is an Arabic numeral, since it is very likely that the locative reference is an address

(18) *South LA 13219 S Penrose Ave **Hit and Run No Injuries***

iii) when one of the preceding tokens is a directional marker

(19) *Accident cleared in #Edmond on NW 178th St at N Pennsylvania Ave #OKCtraffic*

iv) when they are followed by one or a combination of proper nouns, including numbers or directional or movement markers (e.g. *Mount Everest*, *River Thames*)

v) when they are followed by the preposition *of* and one or a combination of proper nouns

(20) *I'm from an upper middle class suburb of Boston.*

No examples of missed locative references were found in relation to the functioning of the rules themselves. It is true, however, that a few went missing because the POS tagger assigned grammatical categories other than nouns for a few location-indicative words in some contexts. In Example (21), *ST* was assigned the adjective POS tag.

(21) *Motor Vehicle Accident - WATERBURY #RT8 South at Exit 34 (WEST MAIN ST #1) at 4/11/2019 10:58:08 AM #cttraffic*

There were a few cases of wrongly retrieved instances, as those in Example (22) and Example (23). In Example (22), *Dr.* was wrongly taken as the abbreviation for the location-indicative noun *drive*, and since the tokens that preceded it were all proper nouns, the whole set of tokens were wrongly considered within the boundaries of a false locative instance. Again, context and a deep-semantic system could have proven essential in disambiguating this type of cases.

(22) *#~~RoadSafetyInitiativeByDSS Saint Dr.~~ MSG has come up with the initiative to tie reflector belts on the stray animals*

In Example (23), *1st church* and *2nd church* were wrongly retrieved by means of the rules that searched for Arabic numerals, which may sometimes be ordinal numbers (e.g. *101th street*).

(23) *@TalbertSwan The ~~1st church~~ burned, everyone thought it could have been an accident. After the ~~2nd church~~ burned, deacons…*

On the other hand, in the case of Spanish and with other Romance languages in mind (i.e. French and Italian), a combination of tokens refers to a locative reference when location-

indicative nouns are followed by one or a combination of proper nouns, sometimes introduced by (a) a preposition, (b) a determiner, or (c) a preposition + determiner, or followed by one number (cf. Flowchart #4 in Appendix 1). The following examples illustrate the locative references extracted on the basis of this rule:

(24) *Incidente vial entre bus ?? ?y un ciclista ????? en la Av. Boyacá con Calle 12, sentido norte- sur. Unidad de ?? @TransitoBta y ?? asignada.*
'Road incident between bus and a cyclist in the Boyacá Ave with 12 Street, northbound-southbound. @TransitoBta unit assigned.'

(25) *#26Ago Accidente vial de camionetas del Sebin en la carretera Higuerote-Curiepe dejó un fallecido.*
'#26Aug Road incident between Sebin vans in the Curie-Higuerote road kills one person.'

(26) *INUNDACIONES EN LA M-40. Imagen de la cámara de la M-40 en el barrio de La Fortuna, en el kilómetro 30.*
'FLOODS IN M-40. Picture from the M-40 camera in the La Fortuna neighborhood, in kilometer 30.'

(27) *La peor parada: inundaciones en Baños de Río Tobía por las tormentas*
'Worst off: floods in Tobía River Baths caused by storms'

(28) *Patrulla de vialidad permanente y campaña concientización, después de accidente en carretera a Boquilla*
'Ongoing road management patrol and awareness campaign after accident in the road to Boquilla'

However, there were a few examples of missed locative references. In Example (29), only *provincias de Ávila* could be extracted, because the regex-based rules could not capture the coordinated items in the NP. Since the number of coordinating items is subject to variation, it is hard to formalize a general pattern without finding exceptions to the rule.

(29) *Inundaciones en las provincias de Ávila, Segovia y Valladolid*
'Floods in the provinces of Ávila, Segovia, and Valladolid'

In Example (30), *Calzada* is not in the location-indicative noun dataset, because it was not subsumed by any of the synsets extracted from EuroWordNet, so the rules could not detect the locative reference.

(30) *Vecinos de #Naucalpan se manifiestan sobre Calzada San Agustín para exigir reforzamiento de muros del Río Hondo*
'#Naucalpan residents protest over San Agustín road to demand the reinforcement of walls in Hondo river'

Moreover, symbols such as the dash, which might occur within the boundaries of locative references, as in Example (31), are not currently dealt with by the rules because these could appear in any position, making the formalization of patterns very hard.

(31) *Alrededor de las 9:10 de esta mañana, una volcadura en la carretera Navojoa - Los Mochis dejó sin vida a una persona.*
'Around 9:10 this morning, rollover in Navojoa – Los Mochis road killed one person.'

In Example (32), the reason why this instance was extracted is due to the fact that *cámara* is a location-indicative noun in the Spanish location-indicative dataset. Since there is not a word-sense disambiguation system in LORE, it is for now impossible to avoid matching ambiguous items whose meaning is different from the location-based one.

(32) *INUNDACIONES EN LA M-40. Imagen de la ~~cámara de la M-40~~ en el barrio de La Fortuna, en el kilómetro 30.*
'FLOODS IN M-40. Picture from the M-40 camera in the La Fortuna neighborhood, in kilometer 30.'

We developed a language-independent rule on the basis of road and highway naming conventions used in English and Spanish-speaking countries, obtained from Wikipedia.[9] If a token includes one or two letters, accompanied or not by the dash symbol, and then followed by a number between 0 and 9999 and an optional letter at the end, then it is very likely that it is the locative reference of a traffic way (i.e. highway or road) (cf. Flowchart #5 in Appendix 1):

(33) *Cortadas por inundación tras la tormenta la M-506, la M-40 y al menos 6 líneas de Metro*
'M-506 and M-40 and at least 6 underground lines blocked because of floods after storm'

(34) *Gracias a la #Tormenta llevamos dos horas parados en la A-42 por inundaciones y sin previsiones de movernos. Genial oye.*
'Thanks to the #Storm we have been kept for two hours in A-43 because of floods, and not expecting to move. Great, huh.'

In English, directional or movement markers may precede or follow highways, which are also captured within the boundaries of the extracted locative references. Moreover, by means of another rule, we account for whitespaces between characters (e.g. *I 84*):

(35) *Update - #M5 northbound J19 #Gordano towards J18 #Avonmouth. Our traffic officers have driven through the area*

(36) *accident:NorthWest Pkwy (TX-114 alt) eastbound TX-26 Grapevine various Lns blocked*

(37) *Incident on #I278 EB from 3rd Avenue to Exit 26 - Hamilton Avenue*

(38) *Motor Vehicle Accident - WATERBURY #RT8 South at Exit 34 (WEST MAIN ST #1) at 4/11/2019 10:58:08 AM #cttraffic*

(39) *One person was killed in an accident on southbound I-91 in New Haven on Thursday morning.*

Example (40) and Example (41) contain locative references missed by these rules.

(40) *Accident on 35W NB @ County Road 96*

In Example (41), the slash symbol, which is not captured by the rules, hampers a successful extraction of the whole locative reference.

[9] For instance, see https://en.wikipedia.org/wiki/List_of_motorways_in_the_United_Kingdom or https://en.wikipedia.org/wiki/Highways_in_Spain

(41) **UPDATE* 15:20?? #M8 E/B J22 Plantation - J18 Charing Cross remains ?CLOSED? due to a police incident on the Kingston.*

Other instances, such as Example (42), Example (43) and Example (44), were wrongly taken as locative references.

(42) *Today #Afghan Army helicopter (~~MD-530~~) crashed dawn due to technical issues while returning from a training operation*
(43) *I have done this by accident and printed tickets ~~A2~~...*
(44) *Terremoto ~~M5.0~~ - Ryukyu Islands, Japan*
'M5.0 earthquake - Ryukyu Islands, Japan'

4.1.4. *Rules that exploit locative markers*

This type of rules can be divided into two main groups: (a) rules that apply to directional markers, and (b) rules that apply to distance and temporal markers (cf. Flowchart #6 in Appendix 1). On the one hand, a combination of tokens containing a directional marker is very likely to refer to a locative reference:

i) when the tokens following the directional marker are proper nouns or locative references previously retrieved, which could be preceded by a preposition (e.g. *de, of*); this rule is language-independent.

(45) *South LA 13219 S Penrose Ave **Hit and Run No Injuries***
(46) *Cleared: Incident on #US9 SB from South of CR 522/Throckmorton St to Exit 26 - Hamilton Avenue*
(47) *Incident on #I78 WB at East of Exit 55 - CR 602/Lyons Ave*
(48) *#VIDEO Este fin de semana, se registraron severas inundaciones al norte de #LosMochis*
'#FOOTAGE This weekend several floods were recorded in the north of #LosMochis'

ii) when the tokens following the directional marker are proper nouns or locative references previously retrieved preceded by a preposition (e.g. *of*), and if the preceding token is a distance marker (e.g. *km*, *miles*) preceded by a number; this rule is English-specific.

(49) *A 3.5 magnitude earthquake occurred 1.86mi SW of San Bernardino, CA.*
(50) *#Earthquake (#tërmet) M2.7 strikes 20 km NW of #Durrës (#Albania) 42 min ago*

iii) when the tokens following the directional marker are proper nouns or locative references previously retrieved preceded by a preposition (e.g. *de*), and the preceding tokens are a number followed by a distance marker (e.g. *kms*, *millas*) followed by a preposition + determiner (e.g. *al*, *del*), e.g. *20 kilómetros al sur de Granada* or *100 millas del suroeste de Londres*; this rule is specific of the Spanish language.

On the other hand, a combination of tokens containing a distance marker (e.g. *km, mile, metro*) or temporal marker (e.g. *horas, hrs, mins*) is very likely to refer to a locative reference:

i) when these markers are preceded by a number and followed by an optional adverb + preposition + optional definite determiner (e.g. *away from, out of, from the, of*) and the following tokens are proper nouns or locative references previously retrieved; this rule is English-specific.

(51) *18:03 Very bad accident just 4 Kms from Narok town*
(52) *Cleared: Motor Vehicle Accident - HARTFORD #I84 West 0.02 miles before Exit 51 (I-91NB) at 4/11/2019 10:56:03 AM*

ii) when these markers are preceded by a number and followed by a preposition + optional definite determiner (e.g. *de, de la, hacia el*) and the following tokens are proper nouns or locative references previously retrieved; this rule is Spanish-specific.

(53) *Se desploma helicóptero matrícula XB-GIL a 6 kilómetros de Tuxtepec, Oaxaca, en la Finca Nuevo Mundo*
'Helicopter with license plate XB-GIL collapses 6 kms from Tuxtepec, Oaxaca, in the Nuevo Mundo Estate'
(54) *A las 22:37 horas, un terremoto de 7.3 grados Richter, con epicentro a 111 km de puerto El Triunfo*
'At 22:37, a 7.6 earthquake, with its epicenter 111 kms from El Triunfo port'

At times, the rules missed or wrongly retrieved locative references. That was the case of Example (55), where *N Iran* was not extracted but only *Iran,* due to the fact that the rules belonging to the location-indicative word module previously extracted *Golestan province N*.

(55) *April 11 -Aqqala, Golestan province N Iran Three weeks after the floods, the houses are still surrounded by floods in Aqqala.*

In Example (56), the coordinated items could not be captured by the rules due to the lack of a formalized pattern for coordination:

(56) *Preocupación por las inundaciones en las zonas este y sur de Madrid, tras la tormenta*
'Worries over floods in the eastern and southern areas of Madrid after storm'

4.2. *Safe-checking rules*

The successful application of the linguistic-based rules must be accompanied by safe-checking rules to ensure that (i) the same extracted locative reference is not repeated, (ii) that boundaries between locative references do not overlap, and (iii) that the boundaries of locative references are well delimited.

In particular, when delimiting the boundaries of locative references, if a detected proper-noun token takes part in another locative reference, either (a) discard the proper-noun token and leave the previously detected locative reference intact, or (b) remove the locative reference, probably wrongly delimited, and add it again with decreased or expanded boundaries. Case (a) applies in all the linguistic processing modules as the last safe-checking rule before adding a potential locative reference that might have already been extracted. For instance, if proper nouns follow a locative preposition, and the first of those was contained in an already-extracted

locative reference from the place-name search in the geodatabase, the safe-checking rule discards those proper nouns. Case (b) is specific to how the linguistic processing module handles location-indicative nouns by expanding the boundaries of previously detected locative references (e.g. *Athens → city of Athens*, *M-30 → autovía M-30*), and also applies to the addition of locative markers to previously detected locative references by expanding their boundaries with these markers (e.g. *Silicon Valley → 40miles SW of Silicon Valley*, *calle Menéndez Pelayo → 15 minutos de la calle Menéndez Pelayo*).

5. CONCLUSIONS

Location extraction plays a critical role in the analysis of microtexts (e.g. tweets) for social sensing, in which agents (e.g. citizens) provide information about their surroundings through social-media services after the interaction with other agents. Applications constructed on social sensors for public security, e.g. emergency management, crime detection, or disease forecasting, and the smart city, e.g. urban administration or intelligent transportation, require a model to identify the locational focus of the event described in the microtext. In this context, fine-grained locative references that take the form of complex linguistic realizations, as in the case of POIs (e.g. *Rockaway Parkway-Canarsie Station*) and traffic ways (e.g. *southbound I-91*), usually remain undetected in standard named-entity recognizers in the field of NLP. The main goal of this article was to describe how such locative references can be automatically detected by the knowledge-based rules in LORE, a proof-of-concept application that exploits linguistic knowledge together with NLP techniques for locative extraction in microtexts. Future research work will focus on finishing the implementation of French and Italian in LORE, while using other NER tools for evaluation, such as Google Entity Recognizer and Stanza. Moreover, to check the usefulness and the generalizability of the performance of our model, we will test LORE with a larger evaluation corpus. We also plan to implement a deep-learning model that feeds off the linguistic knowledge provided by the rules and datasets from LORE to be used as main linguistic-based features, and compare the performance of this model against LORE.

ACKNOWLEDGMENTS

Financial support for this research has been provided by the Spanish Ministry of Science, Innovation and Universities [grant number RTC 2017-6389-5].

REFERENCES

Acheson, E., De Sabbata, S., & Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems, 64*, 309-320. https://doi.org/10.1016/j.compenvurbsys.2017.03.007

Ahlers, D. (2013). Assessment of the accuracy of GeoNames gazetteer data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval - GIR '13* (pp. 74-81). https://doi.org/10.1145/2533888.2533938

Ahmed, M. F., Vanajakshi, L., & Suriyanarayanan, R. (2019). Real-Time Traffic Congestion Information from Tweets Using Supervised and Unsupervised Machine Learning Techniques. *Transportation in Developing Economies, 5*(2). https://doi.org/10.1007/s40890-019-0088-2

Al-Olimat, H. S., Shalin, V. L., Thirunarayan, K., & Sain, J. P. (2019). Towards Geocoding Spatial Expressions. Retrieved from http://arxiv.org/abs/1906.04960

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How Noisy Social Media Text, How Diffrnt Social Media Sources? *International Joint Conference on Natural Language Processing*, (October), 356–364. Retrieved from http://www.aclweb.org/anthology/I13-1041

Barrière, C. (2016). Searching for Named Entities. In *Natural Language Understanding in a Semantic Web Context* (pp. 23-38). Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-41337-2

Bennett, B., & Agarwal, P. (2007). Semantic Categories Underlying the Meaning of 'Place.' In S. Winter, M. Duckham, L. Kulik, & B. Kuipers (Eds.), *Spatial Information Theory* (Vol. 4736, pp. 78-95). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74788-8_6

Bird, S. (2006). NLTK: The Natural Language Toolkit. In *Proceedings of COLING/ACL* (pp. 69-72). Sidney, Australia: Association for Computational Linguistics. https://doi.org/10.3115/1225403.1225421

Cinque, G., & Rizzi, L. (Eds.). (2010). *Mapping Spatial PPs*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195393675.001.0001

Coventry, K. R., & Garrod, S. C. (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Taylor & Francis Routledge. https://doi.org/10.4324/9780203641521

Creary, L. G., Gawron, J. M., & Nerbonne, J. (1989). Reference to locations. In *ACL '89 Proceedings of the 27th annual meeting on Association for Computational Linguistics* (pp. 42-50). https://doi.org/10.3115/981623.981629

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124-147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

Das, R. D., & Purves, R. S. (2019). Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India. *IEEE Transactions on Intelligent Transportation Systems*, 1-10. https://doi.org/10.1109/TITS.2019.2950782

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. In *Expanding the Boundaries of Health Informatics Using Artificial Intelligence: Papers from the AAAI 2013 Workshop* (pp. 20-24). https://doi.org/10.2218/ijdc.v9i1.318

Dutt, R., Hiware, K., Ghosh, A., & Bhaskaran, R. (2018). SAVITR: A System for Real-time Location Extraction from Microblogs during Emergencies. In *WWW'18 Companion: The 2018 Web Conference Companion* (pp. 1643-1649). Lyon, France. https://doi.org/10.1145/3184558.3191623

Eisenstein, J. (2013). What to do about bad language on the internet. *Proceedings of NAACL-HLT 2013*, (June), 359-369. https://doi.org/10.1109/GEOINFORMATICS.2010.5567952

Eke, P. I. (2011). Using Social Media for Research and Public Health Surveillance. *Journal of Dental Research*, *90*(9), 1045-1046. https://doi.org/10.1177/0022034511415277

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, (1995), 363-370. https://doi.org/10.3115/1219840.1219885

Frank, A. U., & Mark, D. M. (1991). Language Issues for Geographical Information Systems. In D. J. Maguire, M. F. Goodchild, & D. W. Rhind (Eds.), *Geographical Information Systems: Principles and Applications* (pp. 147-163). London: Longman Publishers.

Geis, M. L. (1975). English Time and Place Adverbials. *Ohio State University Working Papers in Linguistics*, *18*, 1-11. Retrieved from https://kb.osu.edu/bitstream/handle/1811/81364/WPL_18_June_1975_001.pdf

Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *GeoInformatica*, *17*(4), 635-667. https://doi.org/10.1007/s10707-012-0173-8

Gonzalez-Paule, J. D., Sun, Y., & Moshfeghi, Y. (2019). On fine-grained geolocalisation of tweets and real-time traffic incident detection. *Information Processing and Management*, *56*(3), 1-14. https://doi.org/10.1016/j.ipm.2018.03.011

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, *52*(2), 603-623. https://doi.org/10.1007/s10579-017-9385-8

Hamzei, E., Winter, S., & Tomko, M. (2019). Initial Analysis of Simple Where-Questions and Human-Generated Answers. In S. Timpf, C. Schlieder, M. Kattenbeck, B. Ludwig, & K. Stewart (Eds.), *14th International Conference on Spatial Information Theory (COSIT 2019)* (pp. 1-8). Dagstuhl, Germany: Schloss Dagstuhl: Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.COSIT.2019.12

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., and Luis, V. S., & Javier García Villalba, L. (2019). Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation. *Sensors*, *19*(7), 1-22. https://doi.org/10.3390/s19071746

Herskovits, A. (1985). Semantics and pragmatics of locative expressions. *Cognitive Science*, *9*(3), 341-378. https://doi.org/10.1016/S0364-0213(85)80003-3

Hoang, T. B. N., & Mothe, J. (2018). Location extraction from tweets. *Information Processing and Management*, *54*(2). https://doi.org/10.1016/j.ipm.2017.11.001

Honnibal, M., & Johnson, M. (2015). An Improved Non-monotonic Transition System for

Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1373-1378). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from https://aclweb.org/anthology/D/D15/D15-1162

Hu, Yingjie. (2018a). Geo-text data and data-driven geospatial semantics. *Geography Compass*, *12*(11), 1-19. https://doi.org/10.1111/gec3.12404

Hu, Yingjie. (2018b). Geospatial Semantics. In B. Huang, T. J. Covas, & M.-H. Tsou (Eds.), *Comprehensive Geographic Information Systems* (pp. 80-94). Oxford, UK: Elsevier. https://doi.org/10.1016/B978-0-12-409548-9.09597-X

Hu, Yuheng, Talamadupula, K., & Kambhampati, S. (2013). Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media - ICWSM '13* (pp. 244-253). https://doi.org/10.1.1.297.589

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2014). *Processing Social Media Messages in Mass Emergency: A Survey*. *ACM Computing Surveys* (Vol. 47). https://doi.org/10.1145/3184558.3186242

Ingersoll, G. S., Morton, T. S., & Farris, A. L. (2013). *Taming Text*. Manning Publications.

Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, *22*(3), 219-228. https://doi.org/10.1080/13658810701626343

Jongman, B., Wagemaker, J., Romero, B., & de Perez, E. (2015). Early Flood Detection for Rapid Humanitarian Response: Harnessing Near Real-Time Satellite and Twitter Signals. *ISPRS International Journal of Geo-Information*, *4*(4), 2246-2266. https://doi.org/10.3390/ijgi4042246

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., & Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, *23*(1), 118–136. https://doi.org/10.1111/tgis.12510

Khodabandeh-Shahraki, Z., Fatemi, A., & Tabatabaee Malazi, H. (2019). Evidential fine-grained event localization using Twitter. *Information Processing and Management*, *56*(6), 102045. https://doi.org/10.1016/j.ipm.2019.05.006

Kracht, M. (2002). On the Semantics of Locatives. *Linguistics and Philosophy*, *25*(2), 157-232.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, *16*(2), 217-265. https://doi.org/10.1017/s0140525x00029733

Leidner, J., & Lieberman, M. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, *3*(2), 5-11. https://doi.org/10.1145/2047296.2047298

Levinson, S. C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge, UK: Cambridge University Press.

Li, C., & Sun, A. (2014). Fine-grained location extraction from tweets with temporal awareness. In *SIGIR 2014* (pp. 43-52). https://doi.org/10.1145/2600428.2609582

Liu, F., Vasardani, M., & Baldwin, T. (2014). Automatic Identification of Locative Expressions from Social Media Text. In *Proceedings of the 4th International Workshop on Location and the Web* (pp. 9-16). Shangai. https://doi.org/10.1145/2663713.2664426

Liu, X., Zhou, M., Wei, F., Fu, Z., & Zhou, X. (2012). Joint inference of named entity recognition and normalization for tweets. In *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference* (Vol. 1, pp. 526-535).

Malmasi, S., & Dras, M. (2016). Location mention detection in tweets and microblogs. In K. Hasida & A. Purwarianti (Eds.), *Communications in Computer and Information Science* (Vol. 593, pp. 123-134). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-0515-2_9

Martínez-Rojas, M., Pardo-Ferreira, M. del C., & Rubio-Romero, J. C. (2018). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*, *43*(April), 196-208. https://doi.org/10.1016/j.ijinfomgt.2018.07.008

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location Extraction from Social Media. *ACM Transactions on Information Systems*, *36*(4), 1-27. https://doi.org/10.1145/3202662

Murthy, D. (2018). *Twitter: Social Communication in the Twitter Age* (2nd ed.). Malden, MA: Polity Press.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. In *Foundations and Trends in Information Retrieval* (Vol. 12, pp. 164-318). https://doi.org/10.1561/1500000034

Purves, R. S., & Derungs, C. (2015). From Space to Place: Place-Based Explorations of Text. *International Journal of Humanities and Arts Computing*, *9*(1), 74-94. https://doi.org/10.3366/ijhac.2015.0139

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. New York, USA: Longman Publishers.

Radke, M., Stock, K., & Jones, C. B. (2019). Detecting the Geospatialness of Prepositions from Natural Language Text. In S. Timpf, C. Schlieder, M. Kattenbeck, B. Ludwig, & K. Stewart (Eds.), *14th International Conference on Spatial Information Theory (COSIT 2019)* (pp. 1-8). Dagstuhl, Germany: Schloss Dagstuhl: Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.COSIT.2019.11

Reppen, R. (2010). Building a corpus. In *The Routledge Handbook of Corpus Linguistics* (pp. 31-37). Routledge. https://doi.org/10.4324/9780203856949.ch3

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., … Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. Retrieved from http://arxiv.org/abs/2003.13907

Stock, K., Jones, C. B., & Tenbrink, T. (2019). Speaking of Location: Communicating about Space with Geospatial Natural Language. In K. Stock, C. B. Jones, & T. Tenbrink (Eds.), *Proceedings of the Workshop on Speaking of Location 2019: Communicating about Space co-located with 14th International Conference on Spatial Information Theory (COSIT 2019)* (pp. 1-7). Regensburg, Germany. Retrieved from http://ceur-ws.org/Vol-2455/paper1.pdf

Talmy, L. (2000). How Language Structures Space. In *Toward a Cognitive Semantics* (Vol. 1, pp. 177-254). MIT Press.

Van, T. N., Gaio, M., & Moncla, L. (2013). Topographic subtyping of place named entities: a linguistic approach. In *AGILE 2013* (pp. 1-5). Leuven, Belgium.

Vasardani, M., Winter, S., Richter, K.-F., Stirling, L., & Richter, D. (2013). Spatial interpretations of preposition “at,” 46. https://doi.org/10.1145/2442952.2442961

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., … Anderson, K. M. (2011). Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In *Fifth International AAAI Conference on Weblogs and Social Media* (pp. 385-392).

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI ’10*, (May 2014), 1079. https://doi.org/10.1145/1753326.1753486

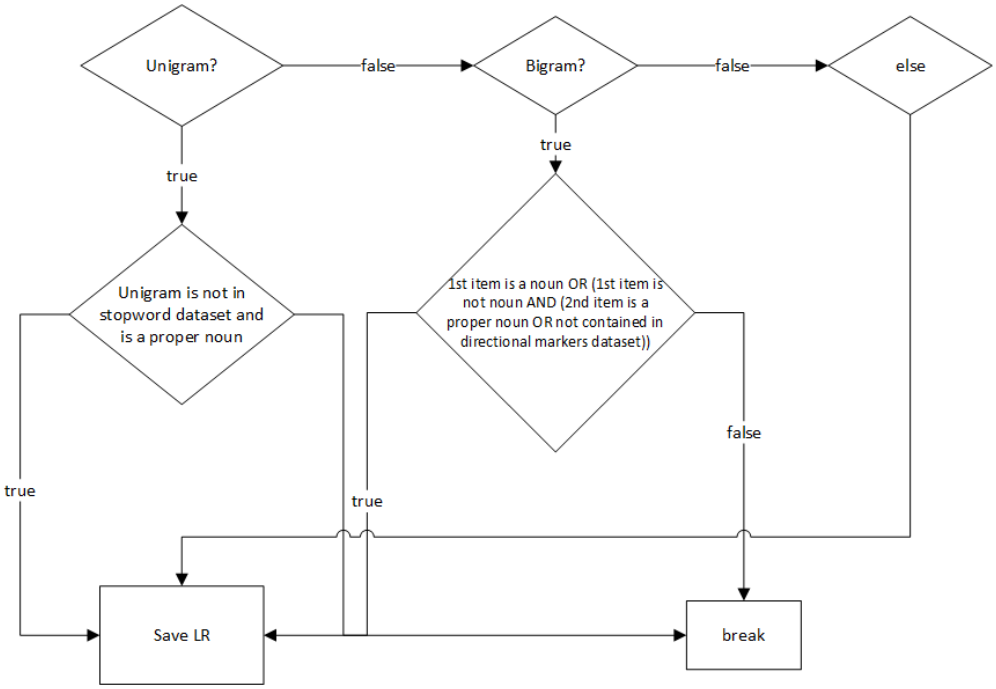
Vilain, P., Menudier, L., & Filleul, L. (2019). Twitter: a complementary tool to monitor seasonal influenza epidemic in France? *Online Journal of Public Health Informatics*, *11*(1), 2017-2020. https://doi.org/10.5210/ojphi.v11i1.9724

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, *32*(1), 1-29. https://doi.org/10.1080/13658816.2017.1368523

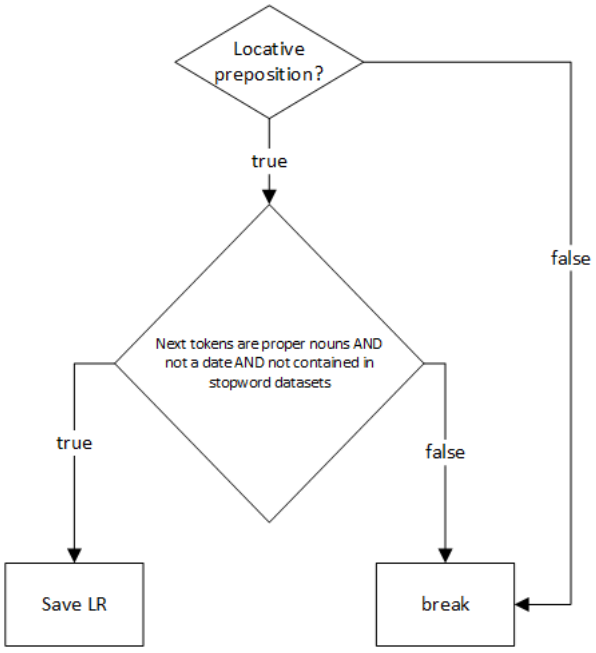
Wang, J., Hu, Y., & Joseph, K. (2020). NeuroTPR : A Neuro-net ToPonym Recognition Model for Extracting Locations from Social Media Messages. *Transactions in GIS*, 1–22.
Zhang, C., Fan, C., Yao, W., Hu, X., & Mostafavi, A. (2019). Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management*, *49*(April), 190-207. https://doi.org/10.1016/j.ijinfomgt.2019.04.004

Appendix 1. Flowcharts of LORE linguistic rules

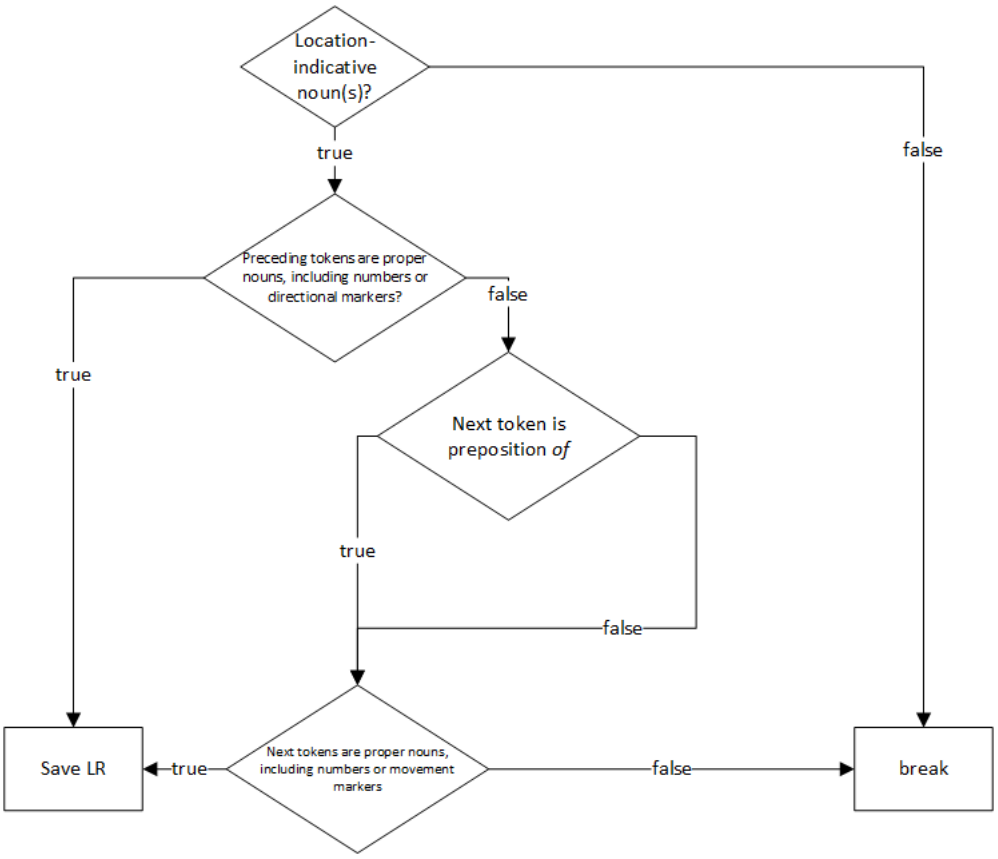
Flowchart #1



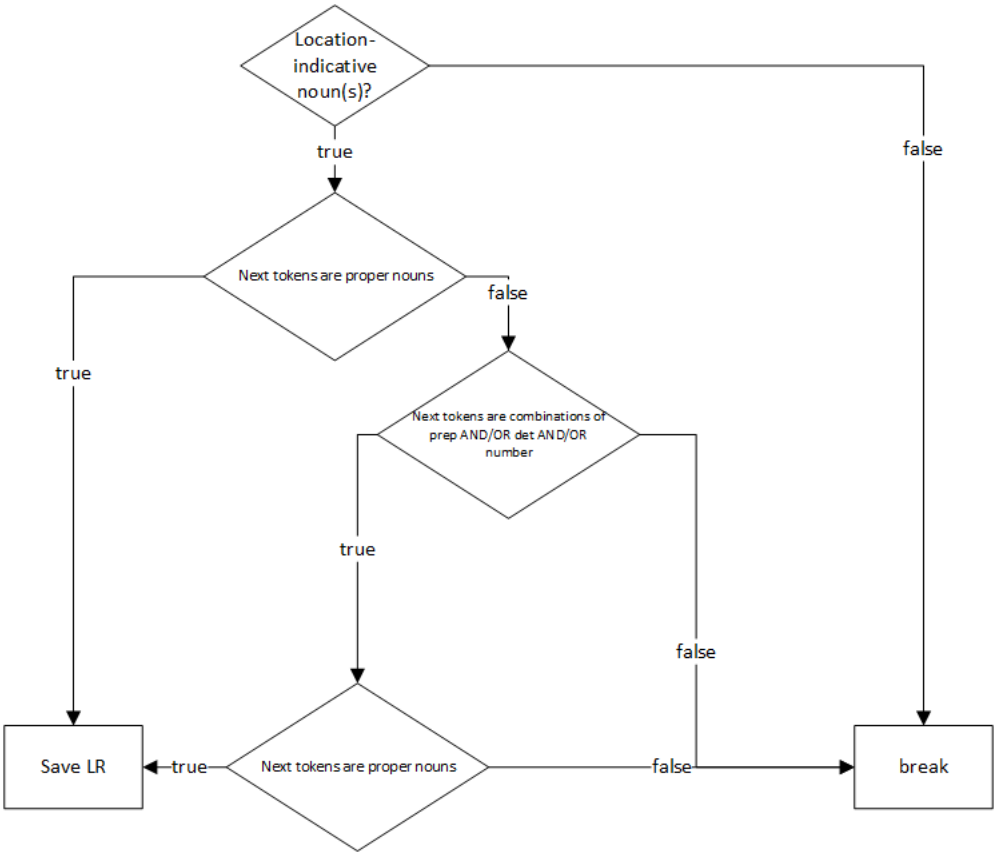
Flowchart #2



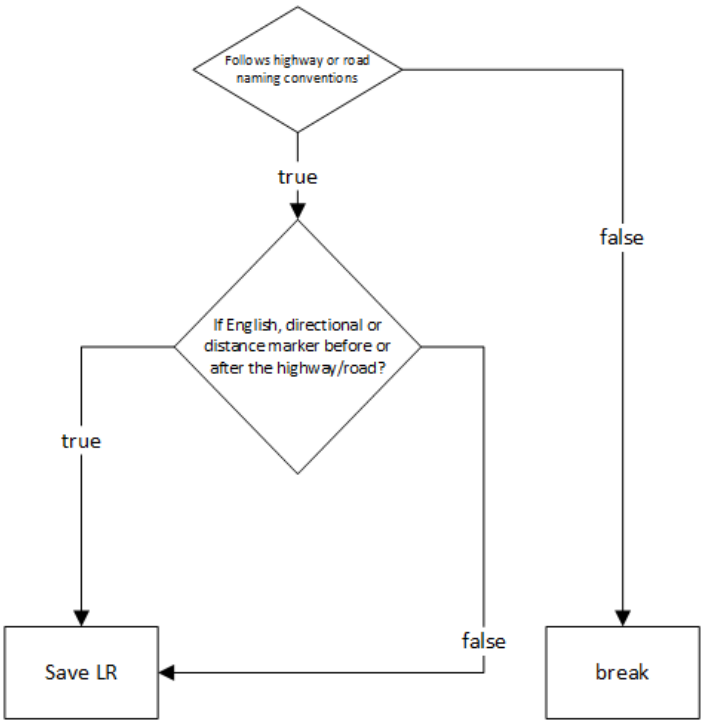
Flowchart #3



Flowchart #4



Flowchart #5



Flowchart #6